Parallel Driving for Fast Quantum Computing Under Speed Limits

Evan McKinney[†], Chao Zhou[§], Mingkang Xia[§], Michael Hatridge[§], Alex K. Jones[†]

[†]Department of Electrical and Computer Engineering, [§]Department of Physics and Astronomy

University of Pittsburgh

Abstract—Increasing quantum circuit fidelity requires an efficient instruction set to avoid errors from decoherence. The choice of a two-qubit (2Q) hardware basis gate depends on a quantum modulator's native Hamiltonian interactions and applied control drives. In this paper, we propose a collaborative design approach to select the best ratio of drive parameters that determine the *best basis gate* for a particular modulator. This requires considering the *theoretical computing power of the gate* along with the *practical speed limit* of that gate, given the modulator drive parameters. The practical speed limit arises from the couplers' tolerance for strong driving when one or more pumps is applied, for which some combinations can result in higher overall speed limits than others. Moreover, as this 2Q basis gate is typically applied multiple times in succession, interleaved by 1Q gates applied directly to the qubits, the speed of the 1Q gates can become a limiting factor for the quantum circuit. We propose a *parallel-drive* approach that drives the modulator and qubits simultaneously, allowing a richer capability of the 2Q basis gate and in some cases for this 1Q drive time to be absorbed entirely into the 2Q operation. This allows increasingly short duration 2Q gates while mitigating a significant source of overhead in some quantum systems. On average, this approach can decrease circuit duration by 17.84% and decrease infidelity for random 2Q gates by 10.5% compared to the best basic 2Q gate, $\sqrt{\texttt{iSWAP}}$.

I. INTRODUCTION

Quantum Computers (QCs) leverage quantum superposition and entanglement which, unlike classical computers, allows the QC core computing element, or *qubit*, to conceptually interact with all other qubits, simultaneously. This provides the promise of solving problems that that are intractable for classical computers. However, currently realized QCs are part of the Noisy Intermediate-Scale Quantum (NISQ) era. NISQ machines with more than a hundred qubits can be readily created [1]; however, the qubit interactions remain limited to small neighborhoods and these quantum operations—or quantum *gates*—have limited fidelity. While these "noisy" quantum operations continue to improve, even the best gates typically do not exceed 99.9% fidelity [2]–[5].

Quantum interactions are realized through qubit-qubit coupling. Coupling is possible when there is a physical connection between the qubits and is governed by a *modulator*. These modulators range from simple as capacitive couplings to more elaborate nonlinear circuits [1], [6], [7]. The major source of error in superconducting QC hardware, which is at the heart of machines by IBM and Google, comes from qubit decoherence. Thus, continued improvements in quantum gate capabilities and speeds are required to increase feasible circuit depth.

A critical component to building better quantum circuits is to identify the *best basis gate* that can be realized by the modulator. The reason for selecting a single basis gate is that calibrating gates is an expensive process. Otherwise, one could just calibrate every single possible gate, or at least each gate that is required for a particular quantum workload. Also, gates must to be calibrated independently between each pair of qubits as each pair will require different parameters and frequencies to be addressed uniquely. Moreover, gate calibrations are finicky processes that drift over time, requiring periodic re-calibration [8]. Thus, a single gate calibration or determining multiple gate parameters as a function of the calibrated gate is necessary to make this process tractable [9].

However, the metric for determining the best basis gate may not be clear. A standard metric for determining the quality of a gate is to calculate its Haar score, which is its coverage of all possible gates among two qubits as represented in a 3D space by the Weyl Chamber [10]. While this is a good representation of the computational power of the gate, many quantum algorithms tend to be reduced to the `CNOT` family of gates to complete their computational work [11]. The remainder of the circuit typically requires the non-entangling `SWAP` gate, primarily to move data on the machine's interconnection topology. Thus, a basis gate that best optimizes these two operations of `CNOT` and `SWAP` is also a useful metric.

In this work, we consider parametrically driven interactions, in which far off-resonant drives activate, usually non-rotating, terms in the coupler to power an effective two-body interaction between a pair of qubits [12]–[14]. The drive amplitude(s) determine the gate speed, while also imprinting a phase on the resultant qubit-qubit coupling. The most prominent examples of these drives are: (1) photon exchange/swapping/conversion between the modes, or *conversion*, and (2) pair-photon creation/annihilation, also called two-mode squeezing, which produces *gain* in parametric amplifiers. These drives can be applied in tandem in a single coupler [15].

One way to identify the set of basis gates enabled by this process is to explore the parametrically driven coupler in the form of its Hamiltonian expression and the drive parameters that can be used to tune different 2Q basis gates. For instance, it is relatively straightforward to show that a Hamiltonian with gain and conversion terms that naturally implements the `iSWAP` family of gates *can also directly implement* the `CNOT` and `B` families of gates as well as other more exotic gates, all by changing the ratio of gain and conversion (see Section II

1

below). However, the pulse times to implement these gates is a function of the drive capacity of the modulator.

All of these parametrically-driven gates depend on an actuator/modulator that has an inherent *speed limit*. The source of this speed limit is the physical limit of the drive capacity of the device for which over-driving the modulator can result in drive lines causing heating, instability in non-linear objects, "bright-state"-ing, bifurcation, chaos, population leakage [16], among others. Speed limits are a fundamental property of parametric couplers, otherwise there would be no fundamental limit to how often every gate can be made $10\times$ faster. Understanding the various physical mechanisms in determining these speed limits with the goal to improve them is an ongoing research effort both for parametrically driven qubit gates and the related field of parametrically driven amplifiers [17]–[19].

To further complicate the selection of the basis gate, the parametric drive terms, *e.g.*, the gain and conversion terms, both contribute towards the speed limit, but combine in a non-linear way. Thus, finding the fastest basis gate can become an optimization function of both the *theoretical computing power* of the gate and the *pulse time* of that gate due to the physical speed limit of that particular ratio of drive parameters.

When conducting decomposition of a quantum circuit, traditionally, a template such that the selected 2Q basis gate is interspersed with 1Q gates (Fig. 1a). Conceptually, the 1Q gates orients the trajectory in a particular direction, then the 2Q basis gate traverses *the Weyl chamber* to a new point. The Weyl chamber is a 3D representation of the possible states of a pair of qubits. Interestingly, to implement a CNOT or SWAP functionality using $\sqrt{\texttt{iSWAP}}$ the first leg in the Weyl chamber is identical, shown as purple. As the point of interest was not yet reached, the direction is re-oriented (1Q gates) before drawing the next line. CNOT is reached in two steps, but the process repeats for SWAP until the point is reached, which happens on the third step. This process is akin to a car driving on established roads. When on a road (path) the car must follow the road, but there are intersections at fixed points where the car can select new roads to follow.

In this paper we propose a collaboratively designed *parallel-drive* technique that drives both the modulator and the qubits directly. Intuitively, if the objective is the fastest path to the destination, stopping to steer adds delay. Instead, by driving both the qubits and the modulator simultaneously the straight paths become curves in the Weyl chamber, which is akin to turning while driving and breaking the restriction of driving on the established road. This is shown by allowing CNOT to be built from iSWAP without intermediate 1Q gates and eliminating one set of interspersed 1Q gates in SWAP (Fig. 1b). Moreover, parallel-drive allows the 2Q gate to increase in Haar volume and in some cases can eliminate the need for interspersed 1Q gates. More, it can expand the 2Q basis set without needing to calibrate new gates individually (assuming there is no cross talk between 1Q and 2Q gates), since there is still only one 2Q basis gate from which the others are created.

Thus, selecting high speed limit gate families and using increasingly short drive pulses while using parallel-drive to
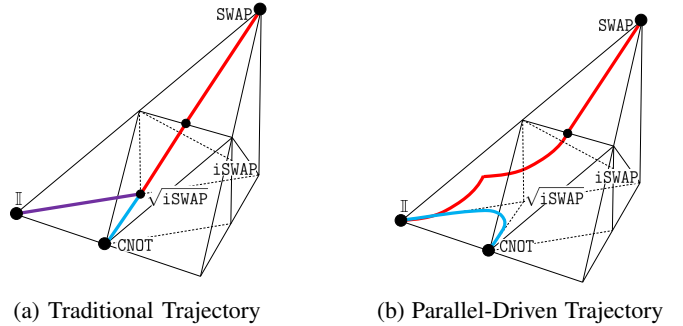

(a) Traditional Trajectory   (b) Parallel-Driven Trajectory

Fig. 1: Cartan Trajectories [20] for CNOT (blue) and SWAP (red) using $\sqrt{\texttt{iSWAP}}$ basis. The trajectories represent the total accumulated unitary transformation over time, beginning at Identity $\mathbb{I}$ and ending at the target gate U_T. Black dots represent interleaved 1Q gates where orientation can be changed.

eliminate some 1Q gates can provide significant speed and fidelity improvements to implemented quantum circuits. In this paper, we make the following contributions:

- We characterize simultaneous application of two basic parametric interactions to implement 2Q gates and articulate the various 2Q gate families that can be realized.
- We demonstrate that these modulators realize biased and potentially non-linear speed limits for different parameter drive ratios.
- We observe that partially pulsed gates, *e.g.*, $\sqrt{\texttt{iSWAP}}$, can be more efficient than the full pulse gate, *e.g.*, iSWAP. However, when using smaller fractions of a gate, the overhead of the 1Q gates becomes more appreciable.
- We present a parallel-drive methodology to improve the agility of a basis gate by concurrently driving the modulator and the participating qubits. We show that parallel-drive can improve the computing capability of a basis gate and provide the potential to remove interleaved 1Q gates in repeated application of 2Q gates.
- Using these approaches we demonstrate an improved equivalency for iSWAP and CNOT by using parallel-drive.
- We present a detailed study of using speed-limits and parallel-drive to reduce circuit delay for important quantum computing workloads.

In the next section we explore the basis gate design space from a modulator by exploring parameters of the Hamiltonian.

II. HAMILTONIAN DESIGN SPACE

Fundamentally, quantum gates are unitary matrix operations, or unitaries, that act on quantum states. In general, 1Q and 2Q gates form the building blocks of quantum circuits [11]. A native quantum gate set, analogous to a classical computer's instruction set, defines which unitary operations are available to use on a machine. The available gates depends on the engineered Hamiltonian of the system, which is related to the unitary, described by Schrödinger's equation, $U(t) = e^{-i\hat{H}t/\hbar}$. In superconducting QCs, parametric driving on a qubit-coupling mechanism provides control over the Hamiltonian to activate the desired unitary and corresponding gate.
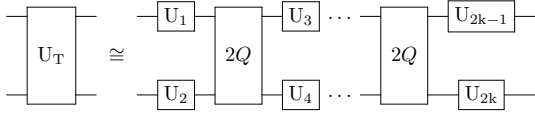
Fig. 2: Generic decomposition 2Q unitary ← 2Q basis gate.

Using Cartan's KAK decomposition [21], [22], an arbitrary 2Q gate can be built from repeated applications of a universal 2Q basis gate with interleaved 1Q gates (Fig. 2). Simple techniques for gate decomposition use this interleaving template and via an exact analytical solution [23], [24], or an approximate numerical optimizer [25]–[27], find a solution to the 1Q gates for a variable number of repetitions. We refer to a *basis template*, as a quantum circuit that interleaves the basis gate K times. To perform decomposition, the template is instantiated with the sufficient size K.

Crucially, the proper selection of basis gate determines the overall complexity of the transpiled quantum algorithm, as different basis gates may require comparatively larger or smaller K in decomposition. Moreover, each basis gate has a latency depending on the system's physical interactions. **For this reason, characterizing the set of candidate basis gates requires reasoning about both their decomposition efficiency, $K[U_B]$, as well as their hardware latency, $D[U_B]$.**

As discussed in Section I, the set of possible 2Q gates is represented geometrically by the Weyl Chamber [28]–[30], where locally-equivalent 2Q gates, differing only by 1Q gates, are mapped to the same coordinate. This comes from the assumption that any *locally equivalent* gate for a particular 2Q gate has the same entangling power and decomposition efficiency. For example, CZ and CX/CNOT can be considered the same equivalent gates in this context. Also, the unitary conjugates are reflected over the x-axis, which is like executing the gate backwards, so essentially it is only necessary to plot gates on the left side of the chamber. In this work, references to a 2Q basis gate may refer generally to the set of locally equivalent gates with matched computational power; however, references to 2Q target gates include the additional 1Q costs of local basis translation required for algorithm correctness.

To reason about decomposition, we plot a gate's coverage volume, which are regions that span all gates buildable by a template. We also use the ∈ notation when referring to a template, which means the 1Q gates are free variables while the 2Q gates are fixed. The use of monodromy polytopes [10] analytically creates the coverage sets, so we can reason about spanning volumes of K gate applications, decide if a gate is contained in a template, and output its weighted volume.

A. Flexible Realization of Gates with Parametric Couplings

While certain classes of gate interactions, such as cross-resonance, rely on linear coupling modes between superconducting qubits [31], [32], there is increasingly widespread use of *parametric* or *tunable* couplers among of groups of two or more qubits, which can be transmons, flux qubits, or high-Q cavities [14], [15], [20], [33], [34]. In these systems,

the coupler's nonlinearity is driven with external flux and/or microwave fields to create a wide array of potential gates. Two well-known families of interactions used in parametric amplification [35] are photon exchange and two-mode squeezing/gain. Photon exchange is produced by driving the coupler to provide the energy source/sink to exhange excitations between the qubits. This can be realized, for instance, in a third-order coupler driven at the qubits' difference frequency.

Two-mode squeezing/gain is produced by driving to provide the energy required for pair production/annihilation in the qubits. In third-order coupling this requires driving at the qubits' sum frequency. Surprisingly, both approaches naturally realize iSWAP gates among two-level or anharmonic qubits. Further, couplings which link to states outside the computational basis that create state dependent phase accumulation (*e.g.*, difference driving between $|11\rangle$ and $|20\rangle$) could be used to produce CZ gates, all in the same system.

These couplers can be driven to produce a wide variety of 2Q gates, especially those in the Weyl chambers' floor, which are just combinations of simultaneous gain and conversion driving[1]. We can write such a combination Eq. 1,

$$\hat{H} = g_c(e^{i\phi_c}a^\dagger b + e^{-i\phi_c}ab^\dagger) + g_g(e^{i\phi_g}ab + e^{-i\phi_g}a^\dagger b^\dagger), \ (1)$$

where g_c, g_g and ϕ_c, ϕ_g represent the pump-controlled amplitude and phase, respectively, such that g_c, ϕ_c result from difference/conversion driving and g_g, ϕ_g are from sum/gain driving. Each choice of control parameters yields a continuum of gates. To illustrate the flexibility of jointly driving multiple interactions simultaneously, the case with both couplings are non-zero strength with both pump phases set to zero arrives at the following unitary:

$$U(t) = \begin{bmatrix} \cos\theta_g & 0 & 0 & -i\sin\theta_g \\ 0 & \cos\theta_c & -i\sin\theta_c & 0 \\ 0 & -i\sin\theta_c & \cos\theta_c & 0 \\ -i\sin\theta_g & 0 & 0 & \cos\theta_g \end{bmatrix} \quad (2)$$

such that $\theta_c = g_c t, \theta_g = g_g t$, where t is the driving time.

By varying the interaction strengths g_c and g_g at a fixed $t = 1$. The iSWAP gate in this language is given by setting θ_c or θ_g to $\frac{\pi}{2}$, yielding Eq. 3.

$$\hat{H} = \frac{\pi}{2}(a^\dagger b + ab^\dagger) \text{ or } \hat{H} = \frac{\pi}{2}(ab + a^\dagger b^\dagger), \quad (3)$$

while the CNOT gate can be realized by setting $\theta_c = \theta_g = \frac{\pi}{2}$, yielding Eq. 4.

$$\hat{H} = \frac{\pi}{4}(a^\dagger b + ab^\dagger) + \frac{\pi}{4}(ab + a^\dagger b^\dagger). \quad (4)$$

There is a continuous set of possible unitary operators that can be naturally realized by this Hamiltonian. By visualizing this in the Weyl Chamber (Fig. 3a) these two points of interest appear at both ends of the yellow band, with the iSWAP at the tip and CNOT along the baseline at the $\frac{\pi}{2}$ point. In fact, the

---

[1]Note, careful attention must be paid to the nonlinearity and encoding of the qubit states being used. For instance, the same parametric interaction among qubits realized as transmons produces different gates than high-Q cavities, for which the latter produces Fock states.
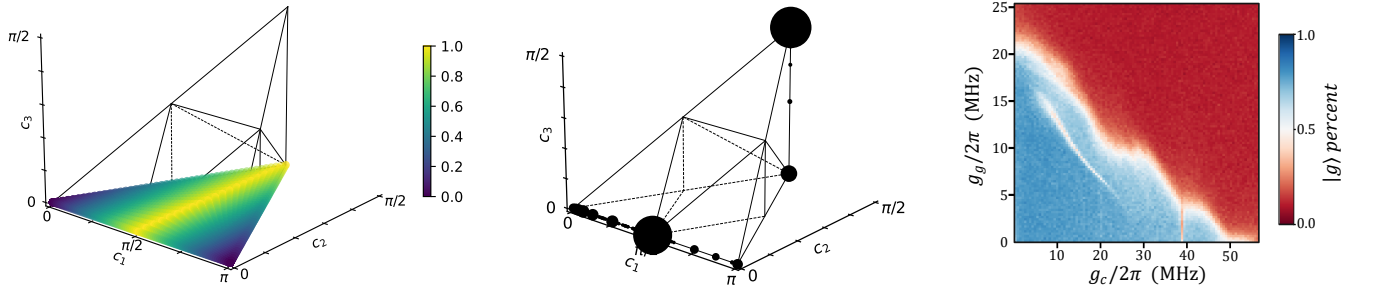
(a) Set of gates natively produced by conversion and gain parametric driving. The color bar indicates $\theta_g + \theta_c$, normalized by $\pi/2$.

(b) Frequency of gates from set of 16-qubit benchmarks transpiled onto an 4×4 square lattice topology.

(c) Demonstration of limitation of gain and conversion coefficients (g_g and g_c) when both processes are turned on.

Fig. 3: Analysis of basis gate choice including, gate range and timing, gate usage by application, and impact of drive ratio.

theoretical power of this Hamiltonian covers the entire base plane of the Weyl chamber with different points reachable in different ratios of θ_g and θ_c and total angle $\theta_g + \theta_c$.

A vital question, then, is which combination of drives yields the best gate? There are several important factors for selecting this gate such as the decompositional efficiency of the gate and the pulse time of the gate. There is evidence that fractional pulse duration gates can be more efficient (*e.g.*, $\sqrt{\texttt{iSWAP}}$ vs $\texttt{iSWAP}$), further reducing pulse time. In the next section we discuss methods to evaluate this decomposition efficiency.

B. Gate Score Methodology

In order to optimize the choice of control parameters, it is first necessary to reason about the unitaries' decomposition efficiency. We compare two methodologies to quantify the decomposition efficiency of a gate: uniform gate distribution and algorithm-sampled distributions. While decomposition determines the number of iterations required of a basis gate to realize a target unitary U_T, recall from Fig. 3a, different realizable gates from the Hamiltonian require different pulse times. To represent both aspects of a gate we define $K_{U_B}[U_T]$ as the number of basis gates (U_B) to build the target (U_T) and $D_{U_B}[U_T]$ as the normalized duration to build a target using the basis. The expectation $\mathbb{E}$ operator signifies the cost averaged over the random Haar distribution.

The Haar measure [36], is used to construct a uniform distribution of 2Q gates. Conceptually, it is a density function inside the Weyl Chamber which weights the perfect entangler interior region more heavily than the exterior $\mathbb{I}$ (identity) and $\texttt{SWAP}$ vertices. It is used to build a Haar score, a common metric to quantify the decomposition power of a basis set. The Haar score is the expected number of gates (K) to generate Haar random 2Q gates. In other words, it is a volume-weighted average over the basis template's spanning regions to achieve full Weyl Chamber coverage. This is demonstrated in Fig. 4, by plotting the K-template spanning region for some popular 2Q gates. The $\texttt{iSWAP}$ gate (Fig. 4a) can reach the bottom plane in $k = 2$ and the entire volume in $k = 3$. The $\sqrt{\texttt{iSWAP}}$ gate (Fig. 4b) actually has better coverage at $k = 2$ with a shorter pulse time. The popular $\texttt{CNOT}$ (Fig. 4c) has similar coverage behavior as $\texttt{iSWAP}$, which is reasonable as both are

(a) $\texttt{iSWAP}$

(b) $\sqrt{\texttt{iSWAP}}$

(c) $\texttt{CNOT}$

(d) $\sqrt{\texttt{CNOT}}$

(e) $\texttt{B}$

(f) $\sqrt{\texttt{B}}$
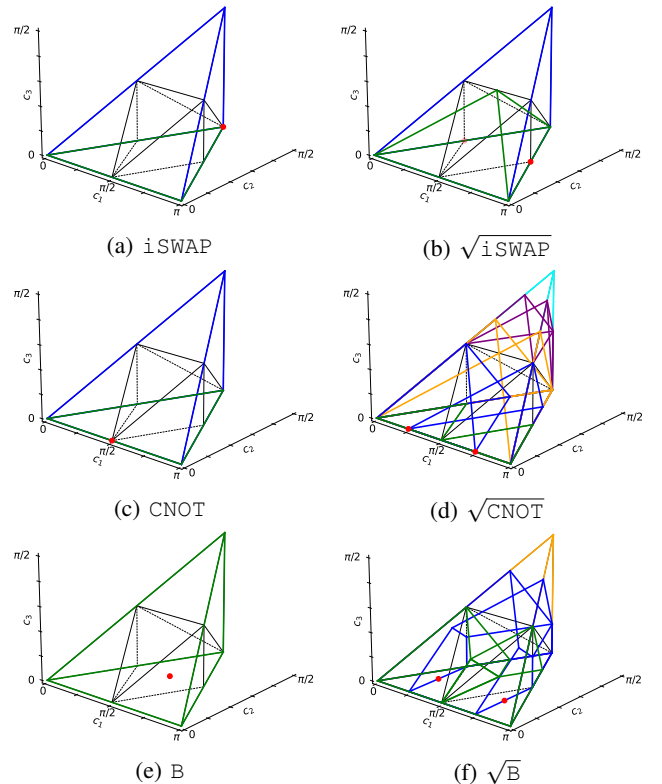
Fig. 4: Gate Coverage Sets. red: $k = 1$, green: $k = 2$, blue: $k = 3$, orange: $k = 4$, purple: $k = 5$, cyan: $k = 6$

Clifford gates. The $\texttt{B}$ gate (Fig. 4e) minimizes $\mathbb{E}[\text{Haar}]$ because it spans the entirety of the region in $k = 2$ (green), whereas $\sqrt{\texttt{CNOT}}$ (Fig. 4d) does not completely span the chamber until six steps ($k = 6$, yellow).

However, the $\mathbb{E}[\text{Haar}]$ score fails to capture that in practice, gates are not uniformly distributed. Algorithms are written primarily using $\texttt{CPhase}$ gates, implemented by *controlled* unitaries, analytically decomposed into 2Q $\texttt{CNOT}$ gates. The reason why $\texttt{CPhase}$ gates are ubiquitous in algorithm design may be explained by the Quantum Singular Value Transform (QSVT), a key subroutine of Grover's Search, Phase Estimation, and Hamiltonian Simulation circuits [37]. The QSVT subroutine encodes an operator $\boldsymbol{A}$ as a block inside a larger unitary matrix, $\boldsymbol{U}$. When $\boldsymbol{A}$ is a unitary matrix,

then U becomes a controlled A operator, which naturally decomposes into `CPhase` 2Q gates [11]. Even algorithms which use different subroutines, such as Quantum Approximate Optimization Algorithm (QAOA), still rely on the `CNOT` for their own reasons. In QAOA, the Hamiltonian cost function maps to states that are diagonal in the computational basis, such that the canonical expansion is into `ZZ` gates [38]. Simply, creating new quantum algorithms is such a difficult task, most known algorithms are variations of the same subroutine, which happens to use controlled-gate operators [39].

Moreover, qubit connectivity topologies necessitate data movement via `SWAP` gates. Due to the limited connection topologies of NISQ superconducting QCs of square lattice and heavy hex, `SWAP` gates are required to move data into qubits in the same neighborhood. It has been shown that these gates can dominate transpiled gate counts [40], [41].

The frequency of `SWAP` gates naturally depends on the coupling topology. For simplicity, we consider a 4x4 square lattice topology as the target coupling map. For a representative set of quantum benchmarks including QFT, QAOA, Adder, Multiplier, GHZ, Hidden Linear Function, and VQE, but excluding the special case of Quantum Volume, the workloads were mapped to this topology using the Qiskit v0.20.2 transpiler with `-O3` (optimization level 3), inducing the necessary `SWAP`s. The results are displayed in a "shot-chart" that increases the size of the plotted gates relative to its frequency in the workloads, as shown in Fig. 3b. From this experiment, the most frequent targeted gates are `SWAP` followed by `CNOT`, with `iSWAP` as a more distant third. Interestingly, there is a significant usage of `CNOT` *family* gates, which show up along the Weyl chamber baseline.

Thus, an alternative gate scoring function introduces $V(U_B)$, which weights the decomposition cost of target, U_T, using the basis gate duration, $D_{U_B}[U_T]$, by the frequency of the target gate, for instance as shown in Fig. 3b. The best basis gate would minimize this weighted cost as shown in Eq. 5.

$$V(U_B) = \sum_{U_T} f(U_T) D_{U_B}[U_T] \qquad (5)$$

As gates must typically be calibrated prior to knowing the circuits they will be programmed to implement, a simplified distribution might only consider and weight the dominating `CNOT` and `SWAP` gates, which, by extension, will generally be true for any `CPhase` algorithm deployed to the device. We fit the value λ as ratio of `CNOT` to the total of `CNOT` and `SWAP` gates using our benchmark workloads as illustrated in Fig. 3b.

$$W(U_B, \lambda) = \lambda * D_{U_B}[\text{CNOT}] + (1 - \lambda) * D_{U_B}[\text{SWAP}] \quad (6)$$

This ratio, $\lambda = 731/(731 + 828) \approx 0.47$. Therefore, the weighted function $W(U_B, 0.47)$ serves to optimize basis gate selection over quantum workload circuits. We go on to refer to this weighted distribution of gates as $W(\lambda = .47)$.

Table I compares the decomposition cost of the six common gates from Fig. 4 in terms of number of gates to realize target gates of `SWAP` and `CNOT`, as well as Haar and our empirical W distributions. The best performing gate for Haar

TABLE I: Decomposition Gate Counts (k). Each value is determined by the spanning regions from Fig. 4

	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{\text{B}}$
K[CNOT]	2	2	1	2	2	2
K[SWAP]	3	3	3	6	2	4
$\mathbb{E}$[K[Haar]]	3	2.21	3	3.54	2	2.50
K[$W(.47)$]	2.53	2.53	2.06	4.12	2	3.06

is the `B` because it can span the Weyl chamber in $k = 2$, however, $\sqrt{\text{iSWAP}}$ and $\sqrt{\text{B}}$ perform well with $k = 2.21$ and $k = 2.5$, respectively. The W cost requires a gate that is good at both `CNOT` and `SWAP`. While the K function is useful for reasoning about theoretical computational capabilities, the D function better compares the implementation of these gates as it considers pulse times and their impact on the decomposition cost, which we explore in the next section.

C. Speed-Limit Scaled Duration Costs

Although each of the discussed candidate basis gates and many others are natively produced by conversion/gain Hamiltonians, different combinations of drives require different duration pulse sequences. It follows from Eq. 1 and Eq. 2, that to realize a specific gate with fixed θ_c and θ_g, the interaction strengths $g_{c,g}$ are inversely proportional to time t. For this reason, a unitary is realizable with the shortest duration when the interaction strengths are as a strong as possible.

However, in a real physical system, the effective g_c and g_g coefficients cannot be infinitely large due to physical limitations, which can include factors such as fridge heating or disrupting parametric coupling [42]. The maximum magnitude is specific to the system being used. In general, it can be described with a Speed Limit Function (SLF) which describes the valid operating range for variable drive strengths. The SLF represents the boundary of the regions where the parameters *obey the speed limit* and coupling operates correctly versus where the *speed limit was exceeded* and the unitary gate fails.

1) Characterizing Gate Speed Limits: To illustrate a concrete example of how the speed limit appears in a parametric coupling system and inform our codesign study, we swept the g_g, g_c drive strengths for a Superconducting Nonlinear Asymmetric Inductive eLement (SNAIL) modulator [43]. The gain-, and conversion-only experiments were first performed individually between a qubit and the SNAIL coupler mode to find the maximum g_c when $g_g = 0$ and vice versa. This calibrates the relations between the drive amplitudes and the g coefficients. Then, the pumps were detuned from the on-resonance frequencies (so that the drive affects the SNAIL but we perform no two-qubit gate) and applied simultaneously to the SNAIL coupler at different amplitude combinations. The result of this study is shown in Fig. 3c.

To monitor the speed limit, which manifests as a *break point* of the SNAIL coupler, a second qubit that also couples to the SNAIL mode is used. This second test qubit is prepared in the ground state and measured immediately after the gain and conversion pumps were turned off. Excitation of this second monitoring qubit signals exceeding the speed limit, in which the SNAIL coupler transitions to a (at present poorly

understood) chaotic behavior and creates photons in both itself and coupled modes, illustrated by the red region in the Fig. 3c. The blue region indicates the monitoring qubit remained in the ground state and represents our proxy for all the feasible g_g and g_c combinations that can be used to construct 2Q gates.

As a result, the SLF of interest is illustrated as the boundary between the blue coupling region and the red non-coupling region, shown as the white line. A few characteristics of interest from Fig. 3c: first, g_c can be driven much harder than g_g and second, the SLF is non-linear.

To capture this experimental information for determining the best basis gate, unitaries described by the values g_c, g_g and time t, a gate can be visualized as a line from the origin with the same g_c to g_g ratio that intersects with the SLF to define g_c^{max} and g_g^{max}. The ratio of change in drive strength is accompanied by inverse scaling of t to find t^{min}. This process is described in Algorithm 1.

To demonstrate the impact of SLFs on basis gate selection we consider two representative synthetic functions in addition to the SNAIL modulator experimental data. The first synthetic function is a linear combination akin to combining voltages, $g_c + g_g \leq L$ resulting in an SLF of $g_g = L - g_c$. The second is a squares of driving strengths function like combining power, enumerated as $g_c^2 + g_g^2 \leq L$ with an SLF of $g_g = \sqrt{L - g_c^2}$.

To more easily compare these functions, we have normalized the SLF to eliminate any dependencies on hardware specific gate durations. For our synthetic functions, this means selecting $L = \pi/2$. For the SNAIL characterized data we uniformly scale such that the maximum x- or y-intercept becomes $\pi/2$. In effect, the fastest iSWAP is normalized to $t^{min} = 1$. Now, rather than reporting $D[U_B]$ in units of time, we use units proportional to 1 iSWAP duration, colloquially referred to as *a single pulse*. In the next section we explore how these SLFs impact decomposition.

2) Circuit Decomposition Costs: Integrating the SLF into duration efficiency combines the theoretical and practical aspects of gate counts to predict circuit latency. Speed-limited duration of the same popular basis gates reported previously are contained in Table II. Compared to the theoretical gate counts where CNOT and B both outperformed $\sqrt{\text{iSWAP}}$, the speed analysis explains why, in practice, $\sqrt{\text{iSWAP}}$ becomes

---

Algorithm 1 Scale Gate Scores using Speed Limit Function

Input: *SLF*, $U_B(\theta_c, \theta_g), K_{U_B}[U_T], D[1Q]$
Find the largest g_c and g_g which produces the input U
$\beta \leftarrow \theta_g/\theta_c$
Find intersection of $g_g = \beta g_c$ with *SLF*(g_c) by solving
$\begin{cases} g_g^{max} = \beta g_c^{max} \\ g_g^{max} = SLF(g_c^{max}) \end{cases}$
Scaling time using updated strengths
$t^{min} \leftarrow \theta_c/g_c^{max}$
Scale decomposition cost by duration
$D_{U_B}[U_T] \leftarrow K_{U_B}[U_T] * t^{min} + (K_{U_B}[U_T] + 1) * D[1Q]$
return $D_{U_B}[U_T]$

TABLE II: Decomposition Duration Efficiency. D_{Basis} is the normalized pulse duration for each candidate basis gate based on the SLF. Then, each decomposition score is computed using Table I and Algorithm 1 over different SLFs. Best value reported in blue, worst value reported in red.

Basis	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{\text{B}}$
Linear Speed Limit						
D_{Basis}	**1.00**	**0.50**	**1.00**	**0.50**	**1.00**	**0.5**
D[CNOT]	2.00	1.00	1.00	1.00	2.00	1.00
D[SWAP]	3.00	1.50	3.00	3.00	2.00	2.00
$\mathbb{E}$[D[Haar]]	3.00	1.05	3.00	1.77	2.00	1.25
D[W(.47)]	2.53	1.27	2.06	2.06	2.00	1.53
Squared Speed Limit						
D_{Basis}	**1.00**	**0.50**	**0.71**	**0.35**	**0.79**	**0.40**
D[CNOT]	2.00	1.00	0.71	0.71	1.58	0.79
D[SWAP]	3.00	1.50	2.12	2.12	1.58	1.58
$\mathbb{E}$[D[Haar]]	3.00	1.05	2.12	1.25	1.58	0.99
D[W(.47)]	2.53	1.27	1.46	1.46	1.58	1.21
SNAIL Characterized Speed Limit						
D_{Basis}	**1.00**	**0.50**	**1.80**	**0.90**	**1.40**	**0.70**
D[CNOT]	2.00	1.00	1.78	1.78	2.81	1.41
D[SWAP]	3.00	1.50	5.35	5.35	2.81	2.81
$\mathbb{E}$[Haar]	3.00	1.11	5.35	3.17	2.81	1.76
D[W(.47)]	2.53	1.27	3.67	3.67	2.81	2.15

the more optimized basis gate, as $\sqrt{\text{iSWAP}}$ has lowest consistent pulse cost for Haar score (1.05–1.11) while $\sqrt{\text{B}}$ has a slightly lower pulse score (0.99) for the squared speed limit. Moreover, $\sqrt{\text{iSWAP}}$ also performs well for W (1.27) while $\sqrt{\text{B}}$ slightly improves on the squared speed limit (1.21).

To find the best basis gate for implementing these target unitaries, the speed limit functions are plotted in Fig. 5. A gate family is defined by the ratio between g_c and g_g terms. The ratios for the CNOT (CX) gate family are shown as blue dotted lines from the origin and the B gate family is a red dotted line from the origin. The iSWAP gate family goes along the x-axis (conversion) and y-axis (gain).

From the figure, the best gate to build a CNOT is directly using a CNOT basis for the linear and squared SLFs. However, on the characterized system, the CNOT is a slow gate, thus it is actually faster to realize an iSWAP basis and convert into CNOT gates. The W function for the squared speed limit is between iSWAP and B on the gain side. For the linear and squared functions the remainder of the gates are at iSWAP on the gain side. For the SNAIL modulator all gates are pinned at iSWAP on the conversion side. Recalling Table II, the basis gate can be from the same gate family but depending on the pulse length, can yield significantly different results.

We must keep in mind from decomposition rules (Fig. 2), templates include interleaved 1Q gates. Our results indicate that for negligible 1Q gate duration, the optimal basis gate is much closer to Identity $\mathbb{I}$, than for appreciable 1Q gates, which tend to be much closer to $\sqrt{\text{iSWAP}}$, CNOT, and B. In the next section we discuss the impact of basis gate selection and fractional pulse lengths as impacted by 1Q gates.

D. Interleaving 1Q Gates

There is a common assumption to treat 1Q gates as negligible, as single qubit interactions are simpler to engineer and as

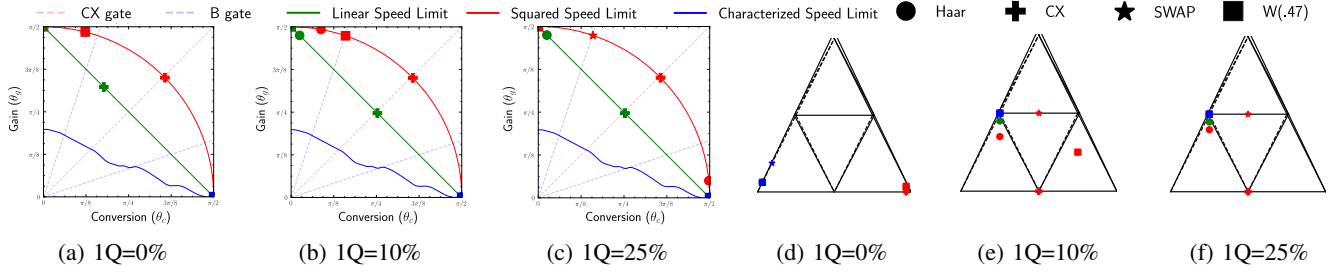| (a) 1Q=0% | (b) 1Q=10% | (c) 1Q=25% | (d) 1Q=0% | (e) 1Q=10% | (f) 1Q=25% |

Fig. 5: Best gates for Haar, CNOT, SWAP metrics for varying 1Q gate durations proportional to the length of the basis gate. (a)–(c) represent a continuum of gates using the ratio between θ_c (x-axis) and θ_g (y-axis), thus the best gate for some metric is where that line intersects with the defined SLF, scaling each gate to the speed limit. (d)–(f) plot gates in the bottom plane of the Weyl chamber to represent the fractional pulse duration. Gates in the bottom left and right corners, near $\mathbb{I}$, are smaller fractions of a pulse versus gates closer to the center, which tends towards durations of a single full pulse.
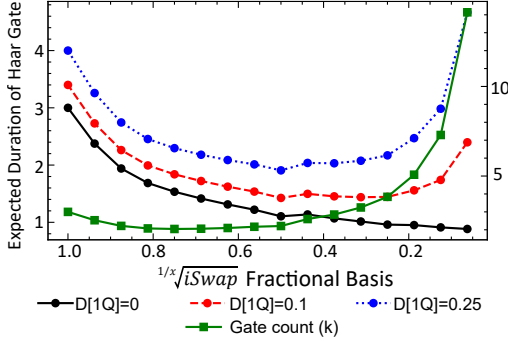

Fig. 6: The optimal Haar gate in the iSwap-family changes as function of the 1Q gate times. This is because the number of gate applications increases for smaller cost 2Q basis gates, which trades off with interleaved 1Q layers. $\sqrt{\texttt{iSWAP}}$ minimizes duration costs for appreciate 1Q gates.

TABLE III: Decomposition duration efficiency. Each value is computed using Eq. 7. ($D[1Q] = 0.25$, Linear SLF)

	iSWAP	$\sqrt{\texttt{iSWAP}}$	CNOT	$\sqrt{\texttt{CNOT}}$	B	$\sqrt{\texttt{B}}$
D[CNOT]	2.75	1.75	1.50	1.75	2.75	1.75
D[SWAP]	4.00	2.50	4.00	4.75	2.75	3.25
$\mathbb{E}$[D[Haar]]	4.00	1.91	4.00	2.91	2.75	2.13
D[W(.47)]	3.41	2.15	2.83	3.34	2.75	2.55

out to near, but not exactly to the iSWAP family.

When considering 1Q gates, the overall duration of a U_B decomposition can be expressed as in Eq. 7, which sums both the 2Q and 1Q durations for K repetitions. To show the impact on decomposition, Table III shows the decomposition efficiency for the linear speed limit when 1Q gates are 25% of the speed of a full pulse 2Q gate. Similar calculations for other speed limits follow the same trends.

$$D_{U_B}[U_T] = K_{U_B}[U_T]t^{min} + (K_{U_B}[U_T] + 1)D[1Q] \quad (7)$$

The total circuit delay can be calculated as in described in Eq. 8, where the pulse delay from Eq. 7 is summed for all gates on the critical path of the full circuit.

$$D_{U_B}[\text{Circuit}] = \sum_{U_T \text{ on Critical Path}} D_{U_B}[U_T] \quad (8)$$

Fig. 6 shows that as the length of 1Q gates increases from 10% of a 2Q gate (red dotted line) to 25% (blue dotted line), the increasingly small iSWAP gates reach a practical limit at $\sqrt{\texttt{iSWAP}}$. This motivates revisiting transpiler optimization targeting 1Q gates. The speed limit formulation indicates that the smaller fractional basis gates are advantageous but only if the increased 1Q gate cost can be mitigated.

Based on the results of this analysis, it is clear that for a linear speed limit, $\sqrt{\texttt{iSWAP}}$ is the most duration optimized basis gate. Furthermore, this methodology offers useful insights for experimentalists when constructing their own basis gates, given their own Hamiltonian design-space and 1Q gate speeds. This is especially pertinent, demonstrated by our hardware speed limit, when there is a strong preference to using one kind of interaction. We also see that as the 1Q gate duration is shorter, the best basis gate approaches Identity, but for appreciable 1Q durations, $\sqrt{\texttt{iSWAP}}$ continues to be optimal.

such, less likely to be a significant source of error. Prior work confirms that $\sqrt{\texttt{iSWAP}}$ is the more efficient basis gate when only considering 2Q gate costs [24], [41]. However, when decoherence is the primary source of error, we find there is an important trade off between faster basis gates and increased K-template lengths (Fig. 6), thus the accumulated 1Q gate count impacts total duration more for fractional basis gates.

In practice, 1Q gates can be quite fast, e.g. around 10% [44] the duration of the basis gate ($D[1Q] = .1$). In other systems with very fast 2Q gates, the 1Q gates are as much as twice as slow as the full pulse 2Q gate [2], [4], depending on the modulator. We treat all 1Q gates as having the same duration, which can be made possible using virtual Z-gates [45].

When the 1Q gate duration is negligible we prefer to have shorter gates with more repetitions. When the speed limit function is convex, we get a better trade off using B or CNOT gates. The best gates tend towards $\sqrt{\texttt{iSWAP}}$ as we observe gates are most efficient when either g_c or g_g is small. Moreover, as shown in Fig. 5, when 1Q gate pulse times are 10% and 25% of a full pulse, respectively, the best basis gates move CNOT to the red dotted line for the linear speed limit and move the W and SWAP to the B family for the squared speed limit at 10% and 25%, respectively. The best Haar score moves

In the next section, we introduce parallel-driven gates as a means of improving the basis gate coverage volumes, and consequently reducing duration costs.

III. PARALLEL 1Q DRIVE FOR BASIS OPTIMIZATION

In the previous section, the power of variable conversion/gain drives was explored to show that a wide variety of basis gates could be realized for different pulse lengths with different optimal driving power ratios determined by the speed limit. However, the same Hamiltonian can be extended if the qubits participating in the 2Q gate are driven simultaneously. In doing so, it is possible to do at least part of the "steering" work of the interleaved 1Q layers during the 2Q gate operation in parallel. This is possible because the drive to the modulator that governs the 2Q interactions is distinct from the drive to the qubits, which would implement the 1Q gates. An important property of this *parallel-drive* methodology is that it is still only necessary to calibrate a single 2Q gate.

A. Parallel-Driven Hamiltonian

Transmon Hamiltonians have been explored to optimize pulses for creation of specific gates or algorithms [46]–[49]. In our work, we modify the Conversion-Gain Hamiltonian by appending single-qubit X-gates with drive amplitudes $\epsilon_1(t), \epsilon_2(t)$ each described by $D[2Q]/D[1Q]$ discrete time steps (Eq 9). Essentially this creates parallel 1Q gates to occupy the duration of the 2Q gate, each with a distinct amplitude.

$$\hat{H} = g_c(e^{i\phi_c}a^\dagger b + e^{-i\phi_c}ab^\dagger) + g_g(e^{i\phi_g}ab + e^{-i\phi_g}a^\dagger b^\dagger)$$
$$+ \epsilon_1(t)(a + a^\dagger) + \epsilon_2(t)(b + b^\dagger) \quad (9)$$

By allowing this extension there are two important outcomes which reduce the overall circuit latency: (1) the basis gate coverage region can be enriched and (2) 1Q gates and their sequential delay may be able to be *absorbed* into the 2Q gate operation, to improve overall circuit time. To consider calibration, parallel-drive applies a frequency Kerr-shift on the qubits, such that the parallel-driven gate would require adjustment for a qubit frequency shift [14], [17]. Note, the SNAIL is a third-order coupler and minimizes the fourth order Hamiltonian term responsible for cross-Kerr interaction. This frequency shift is proportional to the fourth-order term, which will not sacrifice qubit fidelity for the SNAIL. Moreover, it should not impact overhead because existing calibration schemes (*e.g.,* interleaved-randomized and cross-entropy benchmarking) can fine-tune frequency drives of these non-Clifford. Unfortunately, parallel-drive could create additional crosstalk in the IBM cross-resonance gate, however the new IBM initiative to build machines with parametric couplers may also allow high-fidelity parallel-drive. Next, we examine the impact of parallel-drive on Weyl chamber coverage.

B. Computing Parallel Drive Coverage Sets

To demonstrate the additional computing capabilities of basis gates using parallel-driven 1Q gates, we first show the increased set of primitive basis gates ($K = 1$) found by sweeping the free variables of the Hamiltonian (Eq. 9), plotted
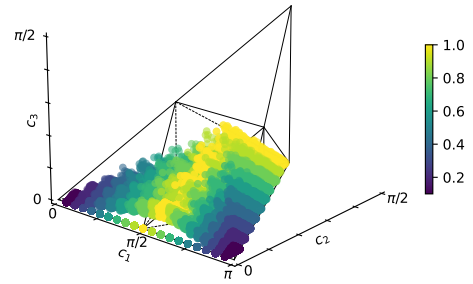

Fig. 7: Gates natively produced by conversion and gain parametric driving with parallel 1Q gate drives ($K = 1$). The color bar indicates the sum of the g_c and g_g, normalized to $\pi/2$.

in Fig. 7 (compared to Fig. 3a). The important outcome is that the parallel-driven basis gates extend off the bottom plane into the volume of the Weyl Chamber, which guarantees that our basis templates with parallel-drive, K', will be able to build some targets with fewer iterations than without, K_0, *e.g.,* $K' \leq K_0$. This translates to an advantage in Haar Score.

As the analytical volume coverage calculation of monodromy polytopes cannot support 2Q gates with parallel-drive, we developed a numerical procedure to estimate the expanded regions. First the coverage region is seeded with randomized parameters. Second, based on the loosely articulated region, target points outside the region are identified to help clarify the outer boundaries. Using an optimizer to these target points a point at the outside edge of the region can be found. The approach is described in Algorithm 2, which constructs a polytope using the `lrs` [50] backend via the convex hulls defined by a set of Weyl chamber coordinates. In order to preserve convexity, we partition the coordinate list into left and right sides of the Weyl chamber ($c_1 = \pi/2$), with convex hulls created separately. Finally, we specifically target exterior points $\mathbb{I}$, `CNOT`, `iSWAP` and `SWAP` as these gates are unlikely to be reached via Haar uniform randomization.

To optimize the template to reach each exterior point, we adapt the strategy from previous work [41], [51], and use the Nelder-Mead optimization method [52] with a Makhlin invariant functional [47], [53]. The free variables are phase ($\phi_{c,g}$) and 1Q drive amplitude ($\epsilon_1(t), \epsilon_2(t)$) bounded by $(0, 2\pi)$, for each K iterations of the template, as shown in Fig. 8a.

We consider four discrete 1Q drive time steps when building the extended volumes. This corresponds $D[1Q] = .25$, for a full pulse `iSWAP`, hence $D[2Q] = 1$. Previous work has explored driving 1Q gates with many more time steps [46]; however, in our experimentation, four time steps provides sufficiently similar coverage sets as compared to 250 time steps, but in a more reasonable computing time. For example, Fig. 8b plots the norm training loss of an `iSWAP` basis converging to `CNOT` in 120 iterations, and Fig. 8c plots the updated coordinate after each iteration. The final (yellow) coordinate successfully converges to the target `CNOT` at $(\pi/2, 0, 0)$. Note that arbitrarily small error is possible with increased training iterations and 1Q time steps.

(a) Decomposition Template (b) Convergence Plot (c) Iterative Gate Coordinate (d) Optimized Cartan Trajectory
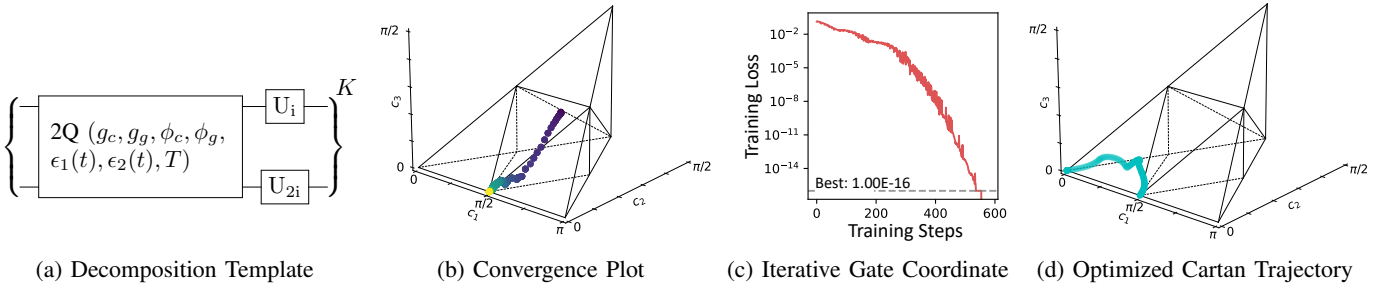
Fig. 8: (a) The decomposition template given to the Nelder-Mead optimizer. To bound the coverage regions we attempt to converge to exterior Weyl chamber points. (b)–(c) $K = 1$ iSWAP is verified to contain CNOT by optimizing over $\epsilon_1(t)$ and $\epsilon_2(t)$. (d) The resulting parallel-drive unitary evolution.

Algorithm 2 Method for Calculating Approximate Improved Volumes from Parallel Drive

---

Basis Template $\leftarrow g_c, g_g, T$
$k \leftarrow 0$
while Coverage Volume not 100% **do**
    $k \leftarrow k + 1$
    Coordinate List $\leftarrow$ []
    Randomly Generate Coverage Points
    for N iterations **do**
        Template $\leftarrow$ Random$(\phi_c, \phi_g, \epsilon_1(t), \epsilon_2(t))$
        $U \leftarrow$ Evaluate(Template)
        $(x,y,z) \leftarrow$ Convert U to Weyl coordinate
        Coordinate List $\leftarrow (x,y,z)$
    end for
    Train for Exterior Coordinates
    for target in (I, CNOT, SWAP, iSWAP) **do**
        Save every coordinate along training path
        Coordinate List $\leftarrow$ Template.optimize(target)
    end for
    Coordinate List partitioned into left and right
    Convex Hulls $\leftarrow$ Coordinate List
    Basis.Polytope[k]$\leftarrow$ Convex Hulls
end while
return Haar Volume(Basis.Polytopes)

C. Impact of Parallel Drive on Decomposition

The extended volumes for each of the six comparative basis gates are reported in Fig. 9. The first major difference from the traditional coverage sets (Fig. 4) is that the $K = 1$ in red has increased from being only local to the basis gate into a non-zero volume. Second, each K spanning region is a superset of its original coverage volumes. Third, no gate reaches 100% coverage in less template repetitions than before, which highlights the inherent difficulty of optimizing the SWAP gate, which is always the last gate to be reached.

Using the same procedure as before, the coverage volumes can be used to find the K and D costs shown in Tables IV and V, respectively. Note, internal 1Q gates are unnecessary if the target gate is identical to multiple fractional copies of the basis gate, *e.g.*, iSWAP formed from two $\sqrt{\text{iSWAP}}$s. While, this seems trivial for traditional circuits where a straight line

(a) iSWAP (b) $\sqrt{\text{iSWAP}}$
(c) CNOT (d) $\sqrt{\text{CNOT}}$
(e) B (f) $\sqrt{\text{B}}$

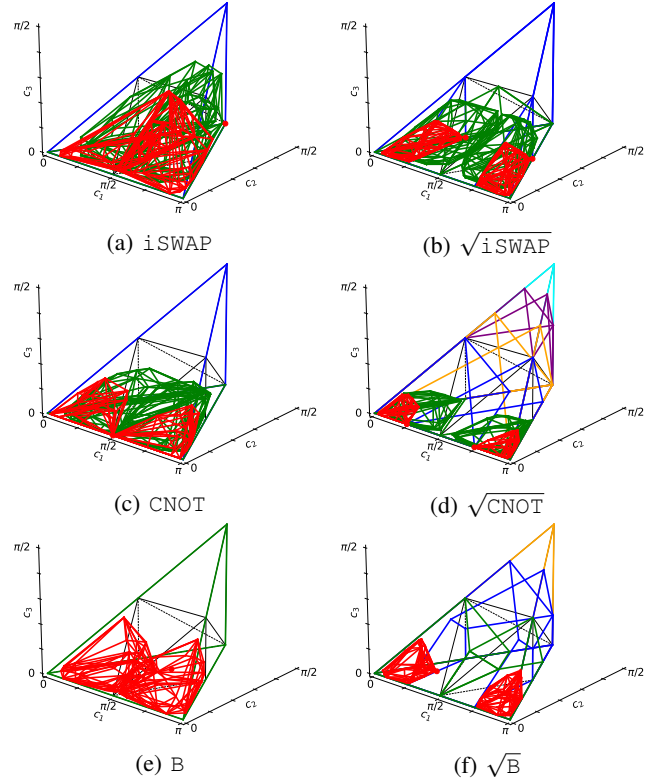Fig. 9: Parallel 1Q Drive Extended Gate Coverage Sets. N=3000 random samples. red: $k = 1$, green: $k = 2$, blue: $k = 3$, orange: $k = 4$, purple: $k = 5$, cyan: $k = 6$

in the Weyl chamber continues twice as far, when adding parallel-drive, this property becomes more important as this line becomes a volume in the Weyl chamber, providing more opportunities to eliminate interleaved 1Q gates.

We use this property to build joint coverage sets between iSWAP and $\sqrt{\text{iSWAP}}$ which create decomposition rules using either gate in the decomposition template. Interestingly, predominately $\sqrt{\text{iSWAP}}$ sees a significant advantage from this procedure, as $K = 1$ iSWAP partially covers the perfect entangling region, which is heavily favored by the Haar distribution. For instance, the $K = 1$ iSWAP volume contains the point $(\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{4})$, which renders the $K = 2$ $\sqrt{\text{iSWAP}}$ coverage of the same point unnecessary, and eliminates the duration from the 1Q gate in the decomposition. Both $\sqrt{\text{CNOT}}$ and $\sqrt{\text{B}}$

TABLE IV: Extended basis decomposition gate count cost (K). Value is determined by the extended spanning regions in Fig. 9

	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{B}$
K[CNOT]	1	2	1	2	1	2
K[SWAP]	2	3	3	6	2	4
$\mathbb{E}[K[\text{Haar}]]$	1.35	2.17	2.33	3.52	1.75	2.50
K[W(.47)]	1.53	2.53	2.06	3.65	1.53	3.06

TABLE V: Extended basis decomposition duration cost (D) using Parallel-Drive, (D[1Q]=.25, Linear SLF). Fractional basis scores are calculated using the joint spanning regions between themselves and the full basis gate, selecting the lowest cost template.

	iSWAP	$\sqrt{\text{iSWAP}}$	CNOT	$\sqrt{\text{CNOT}}$	B	$\sqrt{B}$
D[CNOT]	1.5	1.5	1.5	1.5	1.5	1.5
D[SWAP]	2.75	2.25	4	4	2.75	2.75
$\mathbb{E}[D[\text{Haar}]]$	1.94	1.71	3.16	2.88	2.44	2.06
D[W(.47)]	2.16	1.90	2.83	2.83	2.16	2.16

obey the same rule, but with smaller overlapping volumes. This makes the advantage present, but less significant. After applying parallel-drive to improve the computing power of each basis gate, we continue to find that $\sqrt{\text{iSWAP}}$ is the best candidate for a basis gate. Next, we will build explicit decomposition rules into this basis, implement them into a transpilation scheme and report improved simulated fidelities on quantum algorithm benchmarks.

IV. PARALLEL DRIVE FOR iSWAP-FAMILY

Our work has shown the advantage in calibrating a basis gate with the smallest fraction of total pulse time that does not compromise fidelity. This can improve Haar score by reducing unnecessary computational work done by longer duration gates. However, to support calibration, we propose to set the basis gate to approximately the same pulse duration as a 1Q gate. In the case of the SNAIL modulator, this results in an $\sqrt[4]{\text{iSWAP}}$ such that $D[1Q] = 0.25 = D[2Q]$ Thus the gate can be calibrated without the parallel-drive and again with the parallel-drive to account for the constant frequency shift. This approach minimizes the calibration overhead, as still only a single gate must be calibrated. However, a $\sqrt{\text{iSWAP}}$ and iSWAP can be constructed by two and four $\sqrt[4]{\text{iSWAP}}$s, respectively. Short basis gates are useful for building gates near Identity $\mathbb{I}$, such as the small controlled-phase rotations that appear in QFT; capable of combination to take long strides, *e.g.,* to SWAP; and can take advantage of parallel-drive to boosts the computational power of gates.

A. Parallel Drive for Improving CNOT and SWAP

The methodology of creating coverage sets for parallel-driven gates allows us to easily improve decomposition templates via inspection. The CNOT-family and SWAP decomposition rules are given in Fig. 10 and Fig. 11. For all other gates, the gate coverage set is used as a lookup table for the required template size. Both the CNOT and SWAP decompositions are shown in Fig. 1, where the parallel-drive is responsible for the curve in the trajectory. In a full transpilation
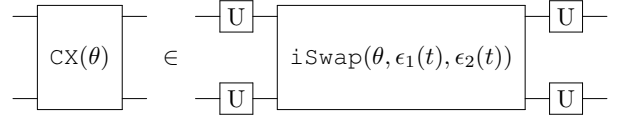

Fig. 10: Decomposition template for CNOT into $\sqrt{\text{iSWAP}}$. A suitable solution for the parallel-drives is $\epsilon_1 = 3, \epsilon_2 = 0$ for all time steps.
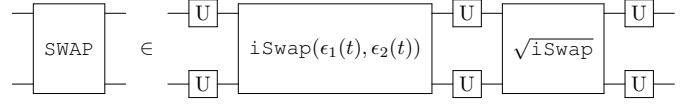

Fig. 11: Decomposition template for SWAP into $\sqrt{\text{iSWAP}}$. A suitable solution for the parallel-drives is $\epsilon_1 = \pi, \epsilon_2 = \pi$ for all time steps. The interior set of 1Q gates is expected to be unnecessary if derived more precisely.
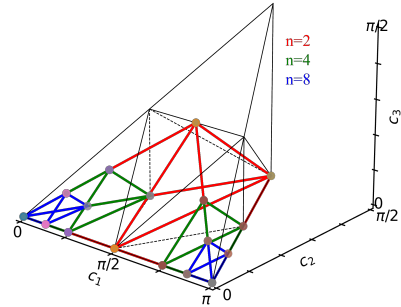

Fig. 12: Illustrating the $K = 2$ coverage of $\sqrt[n]{\text{iSWAP}}$ for $n \in \{2, 4, 8\}$ which can realize $\sqrt[m]{\text{CNOT}}$ for $m \in \{1,2,4\}$.

scheme, the optimizer would be required to fit the exterior 1Q gate parameters, but for the purpose of simulating duration-dependent fidelity, the actual solution is unnecessary.

The Weyl chamber does not represent distances and pulse costs uniformly and may mistakenly convey that a shorter, direct trajectory from $\mathbb{I}$ to U_T reduces the required 2Q basis duration, *e.g.* building CNOT with parallel-driven $\sqrt{\text{iSWAP}}$. However, there is a persistent requirement that 1 total iSWAP pulse durations appear in the decomposition to be able to reach CNOT, likewise 1.5 total iSWAP pulse duration is required to reach SWAP. These inherent costs are more rigorously detailed using quantum resource theories [54], and explain that 2Q decomposition can only be further optimized by removing the 1Q gate delays, but never a shorter 2Q time *i.e.,* the fundamental invariant related to "computing power."

This inherent relationship between iSWAP and CNOT is depicted in Fig. 12, such that a fractional duration iSWAP always contains the same fractional duration CNOT. For instance, K=2 with $\sqrt{\text{iSWAP}}$ or K=1 with parallel-drive iSWAP both reach CNOT. Of course, $\sqrt{\text{iSWAP}}$ is still the more powerful basis despite this relation to CNOT as it always contains additional Weyl Chamber coverage. Both iSWAP and CNOT are special perfect entanglers, and interestingly a non-entangling SWAP gate can be used to convert back and forth between gates [55].

B. Simulated Fidelity Improvements

We utilize a circuit fidelity model that captures the primary source of error as decoherence in time following the methodology of prior work [40], [41]. The fidelity of a final qubit state, $\mathcal{F}_Q$, over a single path or wire in the circuit exponentially decays as a function of the ratio of circuit duration time and the qubit's T_1 (relaxation rate). Then the total circuit fidelity, $\mathcal{F}_T$ is given by the composite final qubit state, and thus is exponential with number of qubits. For this reason, even small improvements in circuit duration cascade into improved path and total circuit fidelities.

$$\mathcal{F}_Q = e^{-D[\text{Circuit}]/T_1} \tag{10}$$

$$\mathcal{F}_T = \prod_{i=1}^{N} \mathcal{F}_{Q_i} \tag{11}$$

Our transpilation scheme uses the SLF normalized durations $D[U_B]$, which are converted back into units of time by multiplying by the `iSWAP` duration. To quantify these improvements we choose $D[\text{iSWAP}] = 100\,\text{ns}$ and $D[1Q] = 25\,\text{ns}$ with qubit lifetime $T_1 = 100\,\mu\text{s}$, which is consistent with transmon qubits using a SNAIL modulator [14]. Using these values, the improvements from the reduced duration decompositions over `CNOT`, `SWAP`, and Haar random targets are given in Table VI. The baseline uses previously derived analytical $\sqrt{\text{iSWAP}}$ decomposition rules [24]. Note that although exterior gates are used in the `CNOT` decomposition to make it perfectly equivalent, quantum algorithms often have 1Q gates before and after `CNOT` gates. Therefore, the circuit's 1Q gates and the decomposition substitution's 1Q gates would naturally combine for an even lower cost than represented here.

In our transpilation flow, we start by consolidating runs of all unitary blocks into 2Q gates and inducing `SWAPs` on a 4×4 square-lattice topology[2]. We then decompose each gate into the $\sqrt{\text{iSWAP}}$ basis with the pulse duration calculated by the provided SLF. Decomposition uses predefined substitutions for gates locally equivalent to `CNOT`-family and `SWAP` gates (see Fig. 1b). If a rule is not known, we load the `iSWAP` and $\sqrt{\text{iSWAP}}$ extended coverage sets to construct a minimum size K template. Finally, we consolidate consecutive 1Q gates and report the remaining durations on each path.

Our decomposition improvements using parallel-drive led to an average relative reduction in duration of 17.8% for the set of quantum algorithm workloads, as determined by selecting the best outcome from 10 transpiler runs, reported in Table VII. The Quantum Volume results were further averaged through additional runs due to the random nature of the algorithm.

The average relative reduction in duration is directly related to our improvement method, while the relative path and total fidelity improvements depend on the baseline duration, mock gate durations, and qubit lifetime. Shallower circuits inherently have higher fidelities, and thus, their improvement potential is limited compared to deeper circuits with lower fidelities. For

---

[2]A `CNOT` followed by a `SWAP` on the same qubit pair is equivalent to an `iSWAP` which appears with non-negligible frequency, see Fig. 3b at `iSWAP`.

TABLE VI: Improved gate infidelities, $1 - \mathcal{F}_Q$ (D[1Q]=.25, Linear SLF)

U_T	Baseline	Optimized	% Improved
CNOT	0.0035	0.0030	14.3
SWAP	0.0050	0.0045	9.98
$\mathbb{E}$[Haar]	0.0038	0.0034	10.5
W(.47)	0.0043	0.0038	11.62

TABLE VII: Transpilation Results (D[1Q]=.25, Linear SLF). Baseline and Optimized columns report total circuit duration in D[2Q]=1 normalized units. Duration, $\mathcal{F}_Q$, and $\mathcal{F}_T$ columns are reported as the relative % improvement between the baseline and optimized durations.

Benchmark	Baseline	Optimized	Duration	$\mathcal{F}_Q$	$\mathcal{F}_T$
QV	133.0	118.4	11.22	1.50	27.0
VQE_L	25.75	21.5	16.50	0.43	7.04
GHZ	31.75	27.00	14.96	0.48	7.90
HLF	102.3	88.00	13.94	1.43	25.6
QFT	149.5	120.3	19.53	2.96	59.5
Adder	175.0	144.3	17.57	3.12	63.6
QAOA	197.8	147.8	25.25	5.12	122
VQE_F	333.3	286.8	13.95	4.76	110
Multiplier	1065.25	770.76	27.64	34.2	11000

instance, the Quantum Volume improves from 0.875 to 0.889 in terms of path fidelities (a 1.5% improvement), which, due to its exponential relationship in the number of qubits, results in an increase from 0.119 to 0.151 (a 21% improvement) in total fidelity. In contrast, the shortest VQE_L algorithm path fidelities baseline of 0.975 only improve to 0.979 (0.4% improvement), leading to a total fidelity increase from 0.662 to 0.709 (6.6% improvement). Finally, our W(.47) metric predicts an average 11.6% reduction in duration. Our experimentally demonstrated 17.8% actually outperforms this case due to additional improvements to `CNOT` where the decomposition template's exterior 1Q gates could be merged or eliminated.

V. CONCLUSION

In this paper, we formally characterized the optimal basis gate for a parametric coupler under hardware speed limitations. The results indicate that, despite the $\sqrt{\text{iSWAP}}$ being close to optimal prior to our analysis, it can still be improved by utilizing parallel 1Q gates. This small improvement leads to a notable enhancement in fidelity as the number of qubits increases. Our co-design evaluated uniform Haar gates and circuit-based gate sets, finding that for realistic cost functions such as our experimentally-determined SNAIL-coupler data, the $\sqrt{\text{iSWAP}}$ gate performed the best in nearly all scenarios.

Initially, gate count scores favored the `B` gate, but after considering the cost of direct generation through multiple simultaneous parametric drives, the $\sqrt{\text{iSWAP}}$ gate was the most efficient. The introduction of parallel-drive and related transpilation optimizations reduced the gate duration for most basis gates and improved the pulse time for the $\sqrt{\text{iSWAP}}$ gate The `iSWAP` family was uniquely enhanced through joint parallel-drive extended coverage sets, yielding significant improvements in fidelity due to faster circuit execution.

In future work, we aim to expand our parallel-drive transpilation flow to further enhance compilation strategies for

quantum algorithms and test them on various quantum systems with differing speed limit characterizations and dynamics. More, detailed studies of improvement of parallel-drive volume versus calibration complexity for different quantum machine targets, including studying calibration complexity, while expanding the flexibility to handle continuously variable drive parameters, similarly to optimal-control theory methods are important next steps.

VI. Acknowledgements

VII. Supplementary Material

The code used in this study is available at https://github.com/Pitt-JonesLab/slam_decomposition. This repository includes the scripts to create circuit templates for decomposition, calculate parallel-driven expanded coverage sets, and implement the transpiler.

References

[1] J. Chow, O. Dial, and J. Gambetta, "Ibm quantum breaks the 100-qubit processor barrier," *[Available Online] https://research.ibm.com/blog/127-qubit-quantum-process-or-eagle*, 2021.

[2] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.

[3] S. Li, A. D. Castellano, S. Wang, Y. Wu, M. Gong, Z. Yan, H. Rong, H. Deng, C. Zha, C. Guo *et al.*, "Realisation of high-fidelity nonadiabatic cz gates with superconducting qubits," *npj Quantum Information*, vol. 5, no. 1, pp. 1–7, 2019.

[4] P. Zhao, P. Xu, D. Lan, J. Chu, X. Tan, H. Yu, and Y. Yu, "High-contrast z z interaction using superconducting qubits with opposite-sign anharmonicity," *Physical Review Letters*, vol. 125, no. 20, p. 200503, 2020.

[5] I. N. Moskalenko, I. A. Simakov, N. N. Abramov, A. A. Grigorev, D. O. Moskalev, A. A. Pishchimova, N. S. Smirnov, E. V. Zikiy, I. A. Rodionov, and I. S. Besedin, "High fidelity two-qubit gates on fluxoniums using a tunable coupler," *npj Quantum Information*, vol. 8, no. 1, pp. 1–10, 2022.

[6] J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, "Simple all-microwave entangling gate for fixed-frequency superconducting qubits," *Physical review letters*, vol. 107, no. 8, p. 080502, 2011.

[7] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow, and J. M. Gambetta, "Error mitigation extends the computational reach of a noisy quantum processor," *Nature*, vol. 567, no. 7749, pp. 491–495, 2019.

[8] C. Tornow, N. Kanazawa, W. E. Shanks, and D. J. Egger, "Minimum quantum run-time characterization and calibration via restless measurements with dynamic repetition rates," *Physical Review Applied*, vol. 17, no. 6, p. 064061, 2022.

[9] D. R. Perez, P. Varosy, Z. Li, T. Roy, E. Kapit, and D. Schuster, "Error-divisible two-qubit gates," *arXiv preprint arXiv:2110.11537*, 2021.

[10] E. C. Peterson, G. E. Crooks, and R. S. Smith, "Fixed-depth two-qubit circuits and the monodromy polytope," *Quantum*, vol. 4, p. 247, 2020.

[11] M. A. Nielsen and I. Chuang, *Quantum computation and quantum information*, 10th ed. Cambridge University Press, 2011.

[12] A. Narla, S. Shankar, M. Hatridge, Z. Leghtas, K. M. Sliwa, E. Zalys-Geller, S. O. Mundhada, W. Pfaff, L. Frunzio, R. J. Schoelkopf *et al.*, "Robust concurrent remote entanglement between two superconducting qubits," *Physical Review X*, vol. 6, no. 3, p. 031036, 2016.

[13] N. Leung, Y. Lu, S. Chakram, R. K. Naik, N. Earnest, R. Ma, K. Jacobs, A. N. Cleland, and D. I. Schuster, "Deterministic bidirectional communication and remote entanglement generation between superconducting qubits," *npj Quantum Inf*, vol. 5, no. 1, pp. 1–5, Feb. 2019. [Online]. Available: https://www.nature.com/articles/s41534-019-0128-0

[14] C. Zhou, P. Lu, M. Praquin, T.-C. Chien, R. Kaufman, X. Cao, M. Xia, R. Mong, W. Pfaff, D. Pekker, and M. Hatridge, "A modular quantum computer based on a quantum state router," *arXiv:2109.06848*, 2021.

[15] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, "Realization of high-fidelity cz and zz-free iswap gates with a tunable coupler," *Phys. Rev. X*, vol. 11, p. 021058, Jun 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.11.021058

[16] L. Theis, F. Motzoi, and F. Wilhelm, "Simultaneous gates in frequency-crowded multilevel systems using fast, robust, analytic control shapes," *Physical Review A*, vol. 93, no. 1, p. 012324, 2016.

[17] G. Liu, T.-C. Chien, X. Cao, O. Lanes, E. Alpern, D. Pekker, and M. Hatridge, "Josephson parametric converter saturation and higher order effects," *Applied Physics Letters*, vol. 111, no. 20, p. 202603, 2017.

[18] L. Planat, R. Dassonneville, J. P. Martínez, F. Foroughi, O. Buisson, W. Hasch-Guichard, C. Naud, R. Vijay, K. Murch, and N. Roch, "Understanding the saturation power of josephson parametric amplifiers made from squid arrays," *Physical Review Applied*, vol. 11, no. 3, p. 034014, 2019.

[19] S. Ashhab, F. Yoshihara, T. Fuse, N. Yamamoto, A. Lupascu, and K. Semba, "Speed limits for two-qubit gates with weakly anharmonic qubits," *Physical Review A*, vol. 105, no. 4, p. 042614, 2022.

[20] S. Lin, S. Sussman, C. Duckering, P. S. Mundada, J. M. Baker, R. S. Kumar, A. A. Houck, and F. T. Chong, "Let each quantum bit choose its basis gates," in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2022, pp. 1042–1058. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/MICRO56248.2022.00075

[21] N. Khaneja and S. Glaser, "Cartan decomposition of su (2ˆ n), constructive controllability of spin systems and universal quantum computing," *arXiv preprint quant-ph/0010100*, 2000.

[22] R. R. Tucci, "An introduction to cartan's kak decomposition for qc programmers," *arXiv preprint quant-ph/0507171*, 2005.

[23] T. Kalajdzievski and N. Quesada, "Exact and approximate continuous-variable gate decompositions," *Quantum*, vol. 5, p. 394, 2021.

[24] C. Huang, T. Wang, F. Wu, D. Ding, Q. Ye, L. Kong, F. Zhang, X. Ni, Z. Song, Y. Shi *et al.*, "Quantum instruction set design for performance," *arXiv preprint arXiv:2105.06074*, 2021.

[25] P. Rakyta and Z. Zimborás, "Approaching the theoretical limit in quantum gate decomposition," *Quantum*, vol. 6, p. 710, 2022.

[26] E. Smith, M. G. Davis, J. M. Larson, E. Younis, L. B. Oftelie, W. Lavrijsen, and C. Iancu, "Leap: Scaling numerical optimization based synthesis using an incremental approach," *ACM Transactions on Quantum Computing*, 2021.

[27] L. Madden and A. Simonetto, "Best approximate quantum compiling problems," *ACM Transactions on Quantum Computing*, vol. 3, no. 2, pp. 1–29, 2022.

[28] Y. Makhlin, "Nonlocal properties of two-qubit gates and mixed states, and the optimization of quantum computations," *Quantum Information Processing*, vol. 1, no. 4, pp. 243–252, 2002.

[29] S. Balakrishnan and R. Sankaranarayanan, "Characterizing the geometrical edges of nonlocal two-qubit gates," *Physical Review A*, vol. 79, no. 5, p. 052339, 2009.

[30] J. Zhang, J. Vala, S. Sastry, and K. B. Whaley, "Geometric theory of nonlocal two-qubit operations," *Physical Review A*, vol. 67, no. 4, p. 042313, 2003.

[31] C. Rigetti and M. Devoret, "Fully microwave-tunable universal gates in superconducting qubits with linear couplings and fixed transition frequencies," *Physical Review B*, vol. 81, no. 13, p. 134507, 2010.

[32] M. Malekakhlagh, E. Magesan, and D. C. McKay, "First-principles analysis of cross-resonance gate operation," *Physical Review A*, vol. 102, no. 4, p. 042605, 2020.

[33] T. Roy, Z. Li, E. Kapit, and D. I. Schuster, "Realization of two-qubit quantum algorithms on a programmable superconducting processor," *arXiv preprint arXiv:2211.06523*, 2022.

[34] D. K. Weiss, H. Zhang, C. Ding, Y. Ma, D. I. Schuster, and J. Koch, "Fast high-fidelity gates for galvanically-coupled fluxonium qubits using strong flux modulation," *arXiv preprint arXiv:2207.03971*, 2022.

[35] A. Roy and M. Devoret, "Introduction to parametric amplification of quantum signals with josephson circuits," *Comptes Rendus Physique*, vol. 17, no. 7, pp. 740–755, 2016, quantum microwaves / Micro-ondes quantiques. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1631070516300640

[36] P. Watts, M. O'Connor, and J. Vala, "Metric structure of the space of two-qubit gates, perfect entanglers and quantum control," *Entropy*, vol. 15, no. 6, pp. 1963–1984, 2013.

[37] J. M. Martyn, Z. M. Rossi, A. K. Tan, and I. L. Chuang, "Grand unification of quantum algorithms," *PRX Quantum*, vol. 2, no. 4, p. 040203, 2021.

[38] E. Farhi, J. Goldstone, and S. Gutmann, "A quantum approximate optimization algorithm," *arXiv preprint arXiv:1411.4028*, 2014.

[39] P. W. Shor, "Why haven't more quantum algorithms been found?" *Journal of the ACM (JACM)*, vol. 50, no. 1, pp. 87–90, 2003.

[40] P. Gokhale, T. Tomesh, M. Suchara, and F. T. Chong, "Faster and more reliable quantum swaps via native gates," *arXiv preprint arXiv:2109.13199*, 2021.

[41] E. McKinney, M. Xia, C. Zhou, P. Lu, M. Hatridge, and A. K. Jones, "Co-designed architectures for modular superconducting quantum computers," in *Proceedings of IEEE Symposium on High Performance Computer Architecture (HPCA)*. arXiv, 2023, [preprint appears online] https://arxiv.org/abs/2205.04387.

[42] C. Zhou, P. Lu, D. Weiss, M. Xia, R. Kaufman, P. Patel, B. Mesits, I. Yusuf, M. Mucci, D. Pekker, J. Koch, and M. J. Hatridge, "Understanding the speed limits of parametrically pumped quantum gates," *Bulletin of the American Physical Society*, 2022.

[43] N. Frattini, U. Vool, S. Shankar, A. Narla, K. Sliwa, and M. Devoret, "3-wave mixing Josephson dipole element," *Applied Physics Letters*, vol. 110, no. 22, p. 222603, 2017.

[44] M. Takita, A. D. Córcoles, E. Magesan, B. Abdo, M. Brink, A. Cross, J. M. Chow, and J. M. Gambetta, "Demonstration of weight-four parity measurements in the surface code architecture," *Physical review letters*, vol. 117, no. 21, p. 210505, 2016.

[45] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, "Efficient z gates for quantum computing," *Physical Review A*, vol. 96, no. 2, p. 022330, 2017.

[46] P. Kairys and T. S. Humble, "Efficient quantum gate discovery with optimal control," in *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, 2021, pp. 413–418.

[47] M. H. Goerz, D. Basilewitsch, F. Gago-Encinas, M. G. Krauss, K. P. Horn, D. M. Reich, and C. P. Koch, "Krotov: A Python implementation of Krotov's method for quantum optimal control," *SciPost Phys.*, vol. 7, p. 80, 2019.

[48] X. Wang, E. Barnes, and S. D. Sarma, "Improving the gate fidelity of capacitively coupled spin qubits," *npj Quantum Information*, vol. 1, no. 1, pp. 1–7, 2015.

[49] N. Earnest, C. Tornow, and D. J. Egger, "Pulse-efficient circuit transpilation for quantum applications on cross-resonance-based hardware," *Physical Review Research*, vol. 3, no. 4, p. 043088, 2021.

[50] D. Avis, "Living with lrs," in *Japanese Conference on Discrete and Computational Geometry*. Springer, 1998, pp. 47–56.

[51] L. Lao, P. Murali, M. Martonosi, and D. Browne, "Designing calibration and expressivity-efficient instruction sets for quantum computing," *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, Jun 2021. [Online]. Available: http://dx.doi.org/10.1109/ISCA52012.2021.00071

[52] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.

[53] P. Watts, J. Vala, M. M. Müller, T. Calarco, K. B. Whaley, D. M. Reich, M. H. Goerz, and C. P. Koch, "Optimizing for an arbitrary perfect entangler. i. functionals," *Physical Review A*, vol. 91, no. 6, p. 062306, 2015.

[54] E. Chitambar and G. Gour, "Quantum resource theories," *Reviews of modern physics*, vol. 91, no. 2, p. 025001, 2019.

[55] G. E. Crooks, "Gates, states, and circuits," 2020.